Découvrir et exploiter le web invisible pour la veille stratégique

Accéder à des milliers de ressources cachées de haute qualité

Janvier 2006

White Paper

Avertissement

Ce document a été réalisé par la société Digimind.

Le contenu de ce document est protégé par le droit d'auteur. Il peut être reproduit en partie à la condition de citer Digimind comme auteur et d'indiquer l'adresse http://www.digimind.com. Pour toute information complémentaire, vous pouvez contacter Digimind par mail à l'adresse contact@digimind.com ou par téléphone au 04 76 70 13 78.

Sommaire

LE WEB INVISIBLE : CONCEPTS ET DEFINITION 4

L'internet n'est pas le web 4

Le Web Visible 5

Le Web Invisible : Définition et structure 6

Un Web Invisible, Caché ou Profond ? 14

CARACTERISTIQUES DU WEB INVISIBLE : TAILLE ET QUALITE 16

La taille du web visible et invisible : estimations 16

Contenu et qualité du web invisible 21

COMMENT IDENTIFIER DES SITES DU WEB INVISIBLE 26

Identifier des sources via les répertoires et annuaires spécialisés 26

Rechercher des ressources spécialisées via votre requête 46

Exemple de ressources du Web Invisible 47

DIGIMIND ET LA VEILLE SUR LE WEB INVISIBLE 49

Les spécificités d'une veille automatisée sur le Web Invisible 49

Comparatif des principaux outils de recherche / surveillance du web invisible* 57

Les avantages de l'approche de Digimind pour la surveillance du Web Invisible 58

A PROPOS DES AUTEURS 60

Digimind Consulting 60

Christophe ASSELIN 60

NOTES 61

Pour rechercher des informations et effectuer votre veille sur le web, l'utilisation des seuls moteurs et annuaires généralistes vous privera de l'identification de centaines de milliers de sources…. Pourquoi ?

Parce que des moteurs comme Google, MSN ou Yahoo! Search[1] ou des répertoires tels que Yahoo! Directory[2] ne vous donnent accès qu'à une petite partie (inférieure à 10%) du web, le Web Visible. La technologie de ces moteurs conventionnels ne permet pas d'accéder à une zone immense du web, le Web Invisible, espace beaucoup plus important que le web visible.

Lors d'une navigation en Antarctique pour prélever des échantillons de glace sur des icebergs, si vous vous limitez à leur partie émergée, vous vous privez de la surface immergée, en moyenne 50 fois plus importante.

Sur le web, c'est la même chose ! **Se contenter du web visible pour votre veille revient à ne pas explorer une zone invisible environ 500 fois plus volumineuse, comportant des centaines de milliers de ressources de grande valeur.**

Les ressources du Web Invisible sont en effet en moyenne de plus grande qualité, plus pertinentes que celles du web de surface. Pourquoi ? Parce qu'elles sont élaborées ou validées par des experts, faisant autorité dans leurs domaines.

Ce document vous présente le concept de Web Invisible ainsi que ses caractéristiques. Par ailleurs, il vous explique comment trouver des sites appartenant à ce web "profond".

Enfin, il vous montre pourquoi l'approche choisie par Digimind, à travers ses nouvelles technologies et son service conseil, vous permet de mettre en œuvre un véritable processus de veille industrielle sur les ressources du Web Invisible.

© Digimind

Le Web Invisible : concepts et définition

Pourquoi un web invisible ?

Pour comprendre le web invisible, Il convient d'abord de rappeler ce que sont le web et sa zone visible.

L'INTERNET N'EST PAS LE WEB

Il existe souvent une confusion et un amalgame sémantique entre le web et l'internet. Ce sont pourtant 2 concepts différents. L'internet, ou Net, est un réseau informatique mondial constitué d'un ensemble de réseaux nationaux, régionaux et privés, interconnectés entre eux (INTERconnected NETworks). Il relie des ressources de télécommunication et des ordinateurs serveurs et clients.

Sur ce réseau, vous pouvez accéder à un certain nombre de services via des protocoles de communication spécifiques :

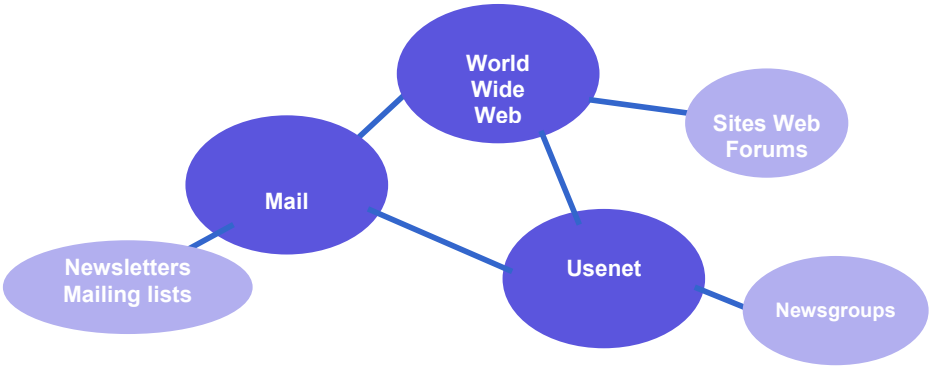
- le World Wide Web (protocole http://www)

- le Mail (smtp)

- Usenet (Newsgroups ou groupes de discussion). Exemple : sci.med.oncology

- le P2P (via des logiciels clients).

et également le FTP, Gopher, …



LE WEB VISIBLE

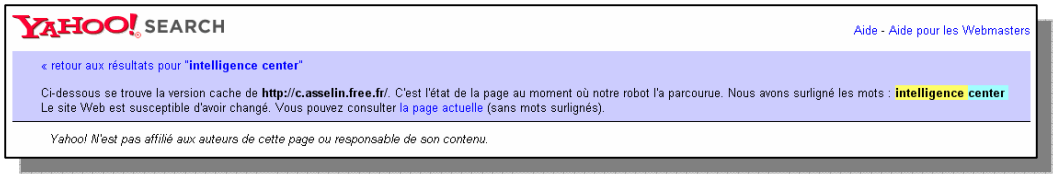
Pour appréhender la notion de Web Visible, il est nécessaire de comprendre le fonctionnement d'un moteur de recherche.

Un moteur de recherche "aspire" puis indexe des pages web. Pour ce faire, ses robots (ou spider) vont parcourir les pages web en naviguant de liens en liens, et passer ainsi d'une page à une autre. Ce robot va archiver sur les serveurs du moteur une version de chaque page web rencontrée.

Ainsi, Google dispose de plus de 10 000 serveurs répartis sur plusieurs états (Californie, Irlande, Suisse…) et installés dans des salles dédiées. Ces serveurs connectés entre eux hébergent notamment les milliards de pages web aspirées.

Cette version archivée des pages au moment du passage du robot est d'ailleurs disponible via la fonction "Cache" disponible sur certains moteurs comme Google, Yahoo!, Clusty ou Gigablast (copies d'écran ci-dessous).

Lors d'une recherche sur un moteur, l'internaute lance une requête sur les bases de données de ces serveurs.



▲ *Les fonctions "Cache" du moteur Yahoo!*

LE WEB INVISIBLE : DEFINITION ET STRUCTURE

Le Web Invisible est constitué des documents web mal ou non indexés par les moteurs de recherche généralistes conventionnels.

En effet, le fonctionnement des moteurs pour "aspirer" le web implique que, d'une part, les pages soient bien liées entre elles via les liens *hypertexte* (http://) qu'elles contiennent et que, d'autre part, elles soient identifiables par les robots du moteur.

Or dans certains cas, ce parcours de liens en liens et cette identification de pages est difficile, voire impossible.

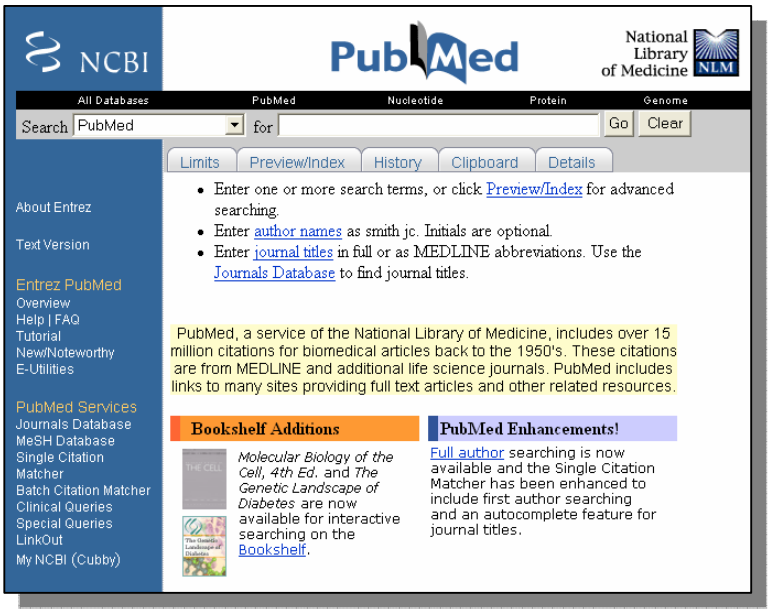
Une partie du web est en effet peu ou non accessible aux moteurs de recherche pour plusieurs raisons :

1- Les documents ou bases de données sont trop volumineux pour être entièrement indexés.

Prenons l'exemple de l'IMDB [3]. L'Internet Movie Database, une base de donnée en libre accès consacrée au cinéma répertorie plus de 7 millions de pages descriptives consacrées aux films et acteurs, représentant chacune une page web. Soit plus de 7 millions de pages. Les moteurs conventionnels n'indexent pas la totalité de ce contenu (son indexation varie entre 5 et 60 % selon les moteurs). C'est aussi le cas de plusieurs milliers de bases de données professionnelles en ligne comme PubMed, certaines n'étant pas indexées du tout.

"Le web invisible est constitué des documents web mal ou non indexés par les moteurs de recherche conventionnels"

D'autre part, les moteurs n'indexent pas la totalité du contenu d'une page lorsque celle-ci est très volumineuse : Google et Yahoo! archivent les pages dans une limite de 500k et 505k.



▲ *La célèbre base PubMed de la National Library of Medecine donnant notamment accès à plus de 15 millions d'articles.*

2- les pages sont protégées par l'auteur (balise meta qui stoppe le robot des moteurs)

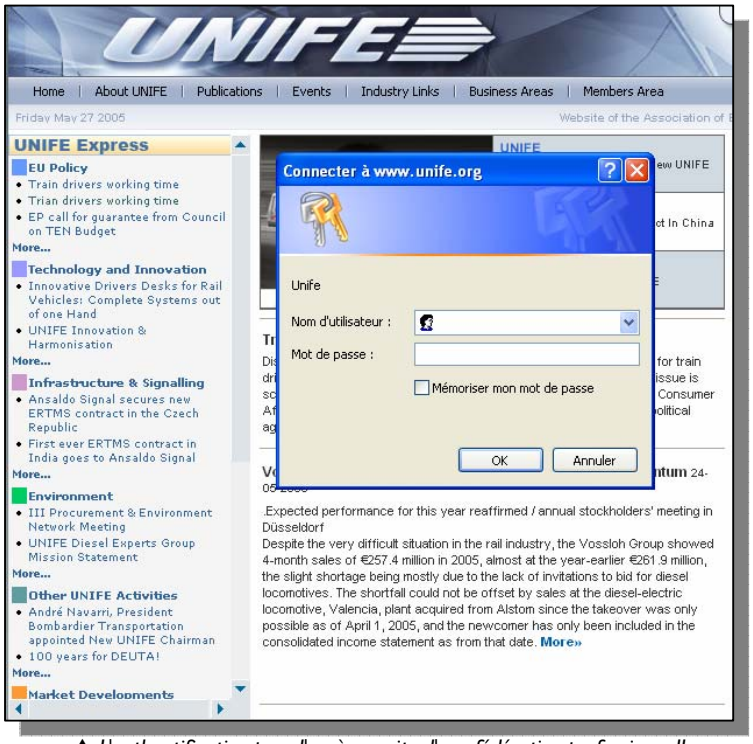
Certains sites sont protégés par leur créateur ou gestionnaire (webmaster), qui, grâce à un fichier *robot.txt* inséré dans le code des pages, interdit leur accès aux robots des moteurs. L'utilisation de ce fichier *robot.txt* est effectuée pour protéger le copyright des pages, limiter leur visite à un groupe restreint d'internautes (auquel on fournira l'adresse, inconnue des moteurs) ou préserver certains sites d'un accès trop fréquent ralentissant les serveurs. Ainsi, le site du journal *Le Monde* interdit aux robots des moteurs de recherche l'accès à ses pages payantes. De cette manière, il évite que les moteurs archivent des pages payantes et les mettent à disposition gratuitement via leur fonction "Cache".

3- les pages sont générées seulement dynamiquement, lors d'une requête par exemple

De nombreux sites web génèrent des pages d*ynamiquement*, c'est-à-dire uniquement en réponse à une requête sur leur moteur interne. Il n'existe pas alors d'URL[4] (adresse) statique des pages que les moteurs pourraient parcourir puisque les robots des moteurs n'ont pas la faculté de taper des requêtes. Ainsi, lorsque vous faites une recherche sur des bases de données ou moteurs, certaines pages générées en réponse à votre question sont structurées avec des *?* et &, de cette manière par exemple : *query.fcgi?CMD=search&*. Ce type de page n'est généralement pas ou mal indexée par les moteurs.

Exemple : *http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=Genome*

4- les pages sont protégées avec une authentification par identifiant (login) et mot de passe. De nombreux sites, qu'ils soient payants ou gratuits, protègent tout ou partie de leur contenu par mot de passe. Les robots de moteurs n'ayant pas la faculté de taper des mots dans des formulaires complexes, ces pages ne leur sont pas accessibles.



▲ *L'authentification pour l'accès au site d'une fédération professionnelle.*

5 - Les formats des documents

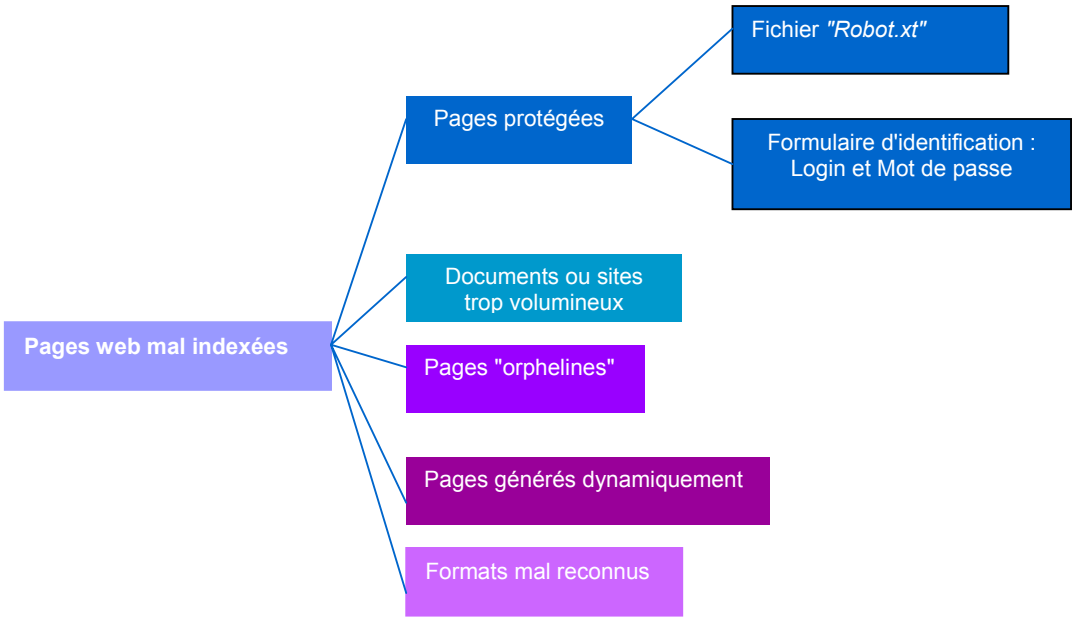
Il y a encore quelques années, on incluait dans le Web Invisible toutes les pages aux formats autres que le html, seul format reconnu et indexé par les moteurs.

En effet, jusqu'en 2001, les moteurs n'indexaient pas les formats PDF (Adobe) et Microsoft Office (Excel, Word, Power Point…). Depuis l'été 2001, Google indexe le PDF et, depuis octobre 2001, il indexe les fichiers Word (.doc), Excel (.xls), Powerpoint (.ppt), Rich Text Format (.RTF), PostScript (.ps) et d'autres encore. Le format Flash, lui, a commencé à être indexé en 2002 par le moteur AlltheWeb [5]

Aujourd'hui, la plupart des grands moteurs à large index (Google, Yahoo!, MSN, Exalead, Gigablast [6]) mémorisent et indexent ces formats de documents. Toutefois, comme nous l'avons vu, l'archivage d'une page web est limité à 500 ou 505 k sur Google et Yahoo! Search. Les documents dépassant ce seuil seront donc partiellement indexés.

On le voit, la structure du web dit "invisible" peut donc évoluer dans le temps en fonction des avancées technologiques des moteurs de recherche. D'"invisibles" avant 2001, certains formats sont devenus "visibles" ensuite.

▼ *Les pages web peu accessibles ou inaccessibles aux moteurs de recherche conventionnels*



6- Les pages sont mal liées entre elles ou sont orphelines (aucun lien présent sur d'autres pages ne pointent vers elles). Plutôt qu'une toile, le web est en fait un immense papillon.

Pourquoi ?

Nous avons vu que le fonctionnement des robots des moteurs de recherche impliquait que toutes les pages soient liées entre elles pour qu'elles soient visibles. Or, ce n'est pas le cas.

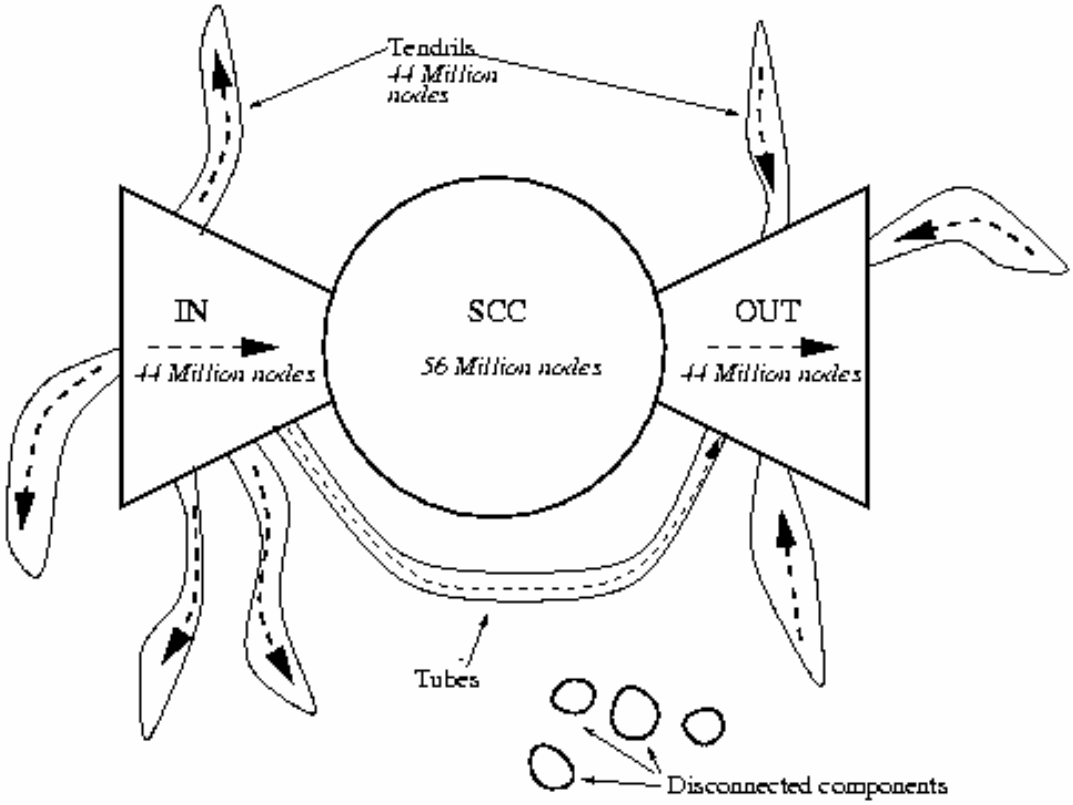
Jusqu'en 2000, on avait coutume de représenter le web à la manière d'une toile d'araignée, supposant ainsi que les pages web sont toutes bien liées entre elles, selon la même densité de liens entrant et sortant, constituant ainsi un ensemble homogène de ramifications. Ainsi, en partant de certaines URLs bien définies, les robots devaient forcément parvenir à parcourir le web en quasi-totalité.

Or, en juin 2000, des chercheurs de IBM (Almaden Research Center, Californie), Compaq et Altavista ont proposé une toute autre représentation du web, plus proche de la réalité…

"Le web, plutôt qu'une toile d'araignée, est structuré comme un immense papillon"

Dans leur étude "*Graph structure in the web*" [7], les 8 chercheurs analysent 200 millions de pages et 1,5 millions de liens collectés par le robot d'Altavista en 3 étapes, de mai à octobre 1999.

L'analyse de ces pages révèle que le web se structurerait en 4 parties :



▲*Connectivity of the web. © Etude "Graph structure in the web"*

1. La partie centrale, le Cœur : *SCC (Giant Strongly Connected Component)*, soit 27,5 % des pages web analysées. Toutes les pages web sont bien connectées entre elles via des liens directs.

2 et 3. Les zones IN et OUT.

La partie IN (21,4%) comprend les pages "source" : ces pages comportent des liens qui pointent vers le cœur (SCC) mais l'inverse n'est pas vrai : aucun lien présent dans le cœur ne pointent sur ces pages. Cela peut concerner de nouvelles pages pas encore découvertes et donc non "liées" par des liens pointant vers elles.

La partie OUT (21,4%) comprend au contraire des pages de "destination" : ces pages sont accessibles et liées à partir du noyau mais en retour, elles n'ont pas de liens qui pointent vers le cœur. C'est le cas, par exemple, de sites "Corporate" de sociétés qui ne comportent que des liens internes et ne proposent pas de liens sortant vers d'autres sites.

4. Les "Tendrils" : (21,4%) : ce terme métaphorique, emprunté à la botanique, désigne des filaments. Ceux-ci représentent des pages qui ne sont pas liées au cœur SCC, et sans liens qui pointent vers elles à partir du cœur.

Ces filaments ne sont connectés qu'aux zones IN et OUT via des liens sortants de la zone IN et des liens entrants dans la zone OUT.

L'étude révèle par ailleurs que si l'on choisit une page au hasard au sein des zones IN et OUT, la probabilité qu'un chemin direct existe de la source (IN) à la destination (OUT) est de seulement 24% (représenté ici par le "tube").

- Enfin, 8,3% des pages constituent des **îlots**, totalement déconnectés du reste du web. Ces pages sont donc orphelines, sans liens pointant vers elles et sans liens sortant, qui auraient permis aux sites destinataires de les trouver. Pour identifier ces pages, il n'existe donc pas d'autres solutions que de connaître leurs adresses fournies par leurs créateurs.

© Digimind

Comment est-il possible que des pages soient mal liées ou non liées ?

Ces pages de sites web peuvent être mal référencées ou non référencées involontairement dans les index de moteurs :

- Par exemple si ses auteurs ne se sont pas préoccupés d'optimiser ces pages (choix adéquat des mots clés, des titres) afin qu'elles soient visibles parmi les premiers résultats des moteurs. C'est ensuite un cercle vicieux : les algorithmes de moteurs comme Google privilégient la popularité des sites pour un bon classement dans les pages de résultats (plus le nombre de liens pointant vers une page est important, plus cette page aura une chance d'être bien classée au sein des résultats). Une page web mal classée dès le départ sera moins visible, donc moins découverte par les internautes et bénéficiera de moins de liens. Ces pages, au mieux, stagneront ou descendront dans les profondeurs des pages de résultats des moteurs.

- Une page placée sur le web sans liens pointant vers elle demeurera invisible des moteurs de recherche. Cela est le cas pour des milliers de documents dont les auteurs ne connaissent pas ou se soucient peu des principes de visibilité de leurs pages (documents de recherche, documents techniques..).

Par ailleurs, on voit sur cette représentation du web en papillon que certaines pages demeurent tout à fait invisibles et d'autres seulement moins visibles …Il existe donc des niveaux de visibilité.

Aussi le terme "*invisible*" au sens strict n'est pas forcément le plus approprié…

© Digimind

UN WEB INVISIBLE, CACHE OU PROFOND ?

Plutôt que de parler d'un Web Invisible, certains spécialistes ont tenté de préciser le concept et de l'illustrer différemment.

Ainsi, Chris Sherman et Gary Price proposent dans leur ouvrage "*The Invisible Web*" [8] de distinguer 4 types de web invisible :

- *The Opaque Web* : les pages qui pourraient être indexées par les moteurs mais qui ne le sont pas (à cause de la limitation d'indexation du nombre de pages d'un site, de leur fréquence d'indexation trop rare, de liens absents vers des pages ne permettant donc pas un crawling de liens en liens)

- *The Private Web* : concerne les pages web disponibles mais volontairement exclues par les webmasters (mot de passe, metatags ou fichiers dans la page pour que le robot du moteur ne l'indexe pas).

- *The Proprietary web* : les pages seulement accessibles pour les personnes qui s'identifient. Le robot ne peut donc pas y accéder.

- *The Truly Invisible Web* : contenu qui ne peut être indexé pour des raisons techniques. Ex: format inconnu par le moteur, pages générées dynamiquement.

Le terme de Web Invisible a été employé pour la première fois en 1994 par le DR Jill Ellsworth pour évoquer une information dont le contenu était invisible des moteurs de recherche conventionnels.

Mais dans son étude "*The Deep Web: Surfacing Hidden Value* " [9], Michael K. Bergman estime, avec raison, que le terme de Web Invisible est inapproprié. En effet, la seule chose d'invisible l'est aux "yeux" des robots des moteurs. Des technologies comme Digimind Evolution permettent d'accéder et de surveiller ce web invisible.

Aussi, plutôt que le web visible et invisible, l'étude de *BrightPlanet* préfère évoquer un "**Surface Web**" et un "**Deep web" (web profond).** En effet, le problème n'est

pas tant la visibilité que l'accessibilité par les moteurs. Il existe donc un web de surface que les moteurs parviennent à indexer et un web profond, que leurs technologies ne parviennent pas encore à explorer, mais qui est visible à partir d'autres technologies perfectionnées.

"Le web n'est pas réellement invisible mais difficilement accessible. Il faut plutôt parler de Web Profond"

On pourrait donc comparer le web à un gigantesque iceberg (en perpétuelle expansion) présentant un volume de ressources immergées beaucoup plus important que les ressources de surface. Comme le web, cet iceberg verrait son volume total augmenter constamment, et la structure de ses parties immergées et émergées se modifier avec le temps (en fonction de l'évolution des technologies des moteurs).

Mais quelle peut être la taille et la composition de ces 2 parties de l'iceberg ?

Si en moyenne, le volume de la partie immergée d'un iceberg est 50 fois plus important que celui de sa partie émergée, les proportions du web visible et invisible sont très différentes…

Caractéristiques du Web Invisible : taille et qualité

Le web profond présente des tailles et caractéristiques bien différentes de celle du web de surface.

LA TAILLE DU WEB VISIBLE ET INVISIBLE : ESTIMATIONS

Plusieurs chercheurs ont tenté d'estimer la taille du web visible puis du web invisible.

Le Web Visible, de Surface

Les études de référence

• Les études de Steve Lawrence et C.Lee Giles du *Nec Research Institute*, Princeton, New Jersey.

Leur première étude, publiée dans le prestigieux *Science Magazine* en avril 1998, "*Searching the World Wide Web*" [10] estime la taille du web de surface (web visible) à 320 millions de documents.

Une mise à jour de l'étude, publiée dans la célèbre revue *Nature* en juillet 1999 "*Accessiblity of Information on the Web*" [11], évalue cette fois le web visible à 800 millions de pages. Cela équivaut alors à une croissance du web de surface de 150 % en 14 mois !

• Début 2000, *NEC*, cette fois en partenariat avec l'éditeur de moteurs *Inktomi* [12] , estime le nombre de pages web à 1 milliard.

"50 millions de pages web seraient modifiées ou ajoutées chaque jour"

• En juillet 2000, une étude de la société *Cyveillance*, "*Sizing the Internet*" [13] estime le web visible à 2,1 milliards de pages, augmentant ainsi à un rythme déjà soutenu de 7,3 millions de pages par jour. Selon ce taux de croissance, Cyveillance estime la taille du web de surface à 3 millions en octobre 2000 et à 4 milliards en février 2001 soit un doublement de volume par rapport à juillet 2000 (ce qui correspond à une période de 8 mois).

Si l'on estime que le web visible double de volume tous les ans, il **pourrait** représenter aujourd'hui, fin 2005, au moins 64 milliards de pages. C'est une estimation évidemment très minorée puisque basée sur un taux de croissance stable ce qui est très peu probable. Ainsi, la possibilité d'accès à la publication sur le web à des utilisateurs toujours de plus en plus nombreux (via les weblogs par exemple qui dépasse les 10 millions) ne fait qu'augmenter son taux de croissance mois après mois.

Capacité des moteurs généralistes actuels

Aujourd'hui, les moteurs généralistes proposant les plus larges index annoncent recenser jusqu'à 8 milliards de pages. A l'été 2005, Google affiche ainsi un "*Nombre de pages Web recensées de 8 168 684 33* ". Yahoo! Search se distingue en août 2005, en annoncant indexer **19, 2 milliards de pages web**.

Or, les chercheurs du centre *IBM d'Almaden* (Californie) estiment que 30% du web est constitué de pages dupliquées et que 50 millions de pages sont modifiées ou ajoutées chaque jour [14] . D'autre part, depuis quelque mois, la taille des index des moteurs évolue peu. Certains estiment même, que du fait de limites technologiques, ils ont atteint leur maximum de taille d'index. Les moteurs actuels n'auraient tout simplement pas la capacité technique d'indexer toutes les pages qui leurs sont accessibles.

Dans ces conditions, un des meilleurs moteurs actuels indexe moins de 8% du web visible. Mais on peut tempérer ce faible taux de représentation du web visible en précisant que l'index de Google n'est pas équivalent à celui de *Yahoo! Search* ou de *Ask Jeeves* [15].

Pour estimer la taille de l'indexation du web visible, il faut donc considérer les index de plusieurs moteurs et non d'un seul.

En effet, les index des moteurs se chevauchent de moins en moins, c'est-à-dire que le nombre de résultats communs à plusieurs moteurs diminue, les résultats uniques proposés par un seul moteur étant en forte progression.

"Les index des moteurs se chevauchent de moins en moins, c'est-à-dire que le nombre de résultats communs à plusieurs moteurs diminue pour atteindre moins de 5% !"

Ainsi, une étude du printemps 2005 réalisée par le métamoteur *Dogpile.com* en collaboration avec des universités de Pennsylvanie estime que les résultats des premières pages fournis par les 3 moteurs à large index que sont *Google, Yahoo! Search* et *Ask Jeeves* sont de plus en plus différents [16].

Sur les premières pages de résultats de chacun de ces moteurs, et sur 336 232 résultats analysés :

- 3% des résultats seulement sont partagés par les 3 moteurs

- 12 % sont partagées par 2 des 3 moteurs

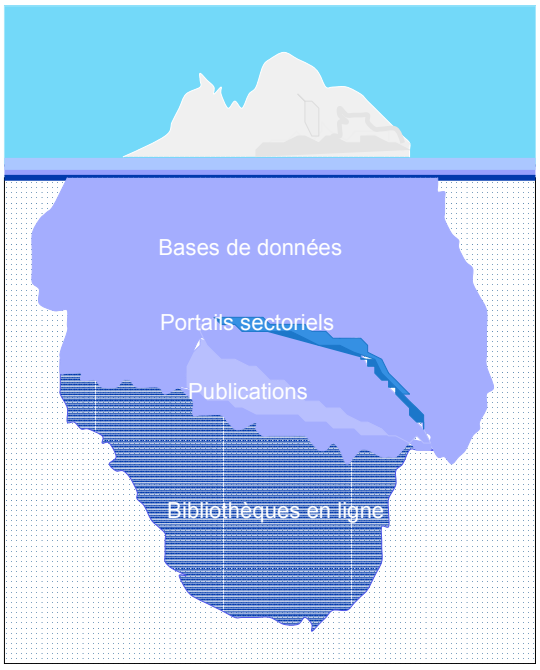
- 85 % ne sont proposés que par un seul des 3 moteurs !

L'étude du métamoteur *Jux2*, sur un échantillon toutefois moins élevé, va également dans ce sens [17]

Un enseignement simple mais riche de conséquence est à tirer de ces études :

Afin de couvrir le mieux possible le web de surface, il faut utiliser **au moins 2 moteurs** pour se rapprocher d'un taux de couverture de 10%. L'utilisation d'un seul moteur généraliste vous fait manquer de nombreux résultats sur le web visible. Or, en France, la pratique des moteurs ne va pas dans ce sens : Google y est prédominant puisque sa part de marché dépasse les 77 % [18], situation unique sur toute la planète. Ainsi, aux USA, si Google est aussi en tête, sa domination est moins forte et les parts de *Yahoo! Search*, *MSN* et *AskJeeves* ne sont pas négligeables.

De par l'utilisation d'un seul moteur généraliste, vos possibilités d'accès au web profond s'en trouvent d'autant réduite. En effet, nous verrons en partie III que certains sites du web de surface peuvent vous aider à accéder au Web Profond.



Le web peut-être représenté par un iceberg (vue en coupe) ▲

La partie émergée symbolise le web visible et la partie immergée, le web profond. Via les moteurs généralistes traditionnels, seule la partie émergée est accessible. Via l'approche de Digimind, la partie immergée sera facilement accessible et ce, via un process d'automatisation permettant une facilité et une industrialisation de la surveillance de ce web profond. *(illustration d'après l'image composite de Ralph A. Clevenger)*

Web invisible, web profond

Les études

C'est l'étude "*The Deep Web: Surfacing Hidden Value*" [8], de Michael K. Bergman qui propose les estimations de la taille du web les plus avancées. Cette étude de 2001 propose des ordres de grandeur permettant de mieux mettre en perspective le web profond à l'égard du web de surface.

En voici les résultats clés :

- l'information publique sur le web profond est considérée comme de **400 à 550 fois plus volumineuse que le web de surface (web visible)**

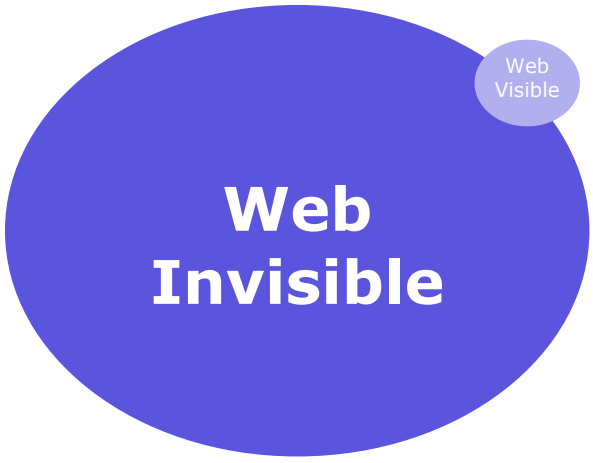
- le web profond est constitué de plus de 200 000 sites web.

- 60 % des sites les plus vastes du web profond représentent à eux seuls un volume qui excède de 40 fois le web de surface.

- le web profond croit plus vite que le web visible.

- plus de la moitié du Web Profond est constitué de Bases de données spécialisées.

- **95% du contenu du web profond est accessible à tous** (gratuit ou à accès non restreint à une profession)



Si l'étude date de 2001, les proportions restent valables (en estimation basse) compte tenu des taux de croissance très élevés du volume du web profond, c'est à dire une multiplication par 9 chaque année (*estimations IDC*).

Les ressources présentes sur le web profond sont donc très nombreuses mais que proposent-elles ?

CONTENU ET QUALITE DU WEB INVISIBLE

Contenu et couverture des sites du web invisible

Typologie du contenu des sites du web invisible

A travers l'analyse de 17000 sites appartenant au web profond, l'étude de *Bright Planet* a élaboré une typologie du contenu du web invisible.

Ces sites ont ainsi été répartis en 12 catégories :

1. Les Base de données spécialisées : il s'agit d'agrégateurs d'informations, interrogeables, spécialisés par sujet. Ce sont par exemple les bases de données médicales, de chimie, de brevets. Exemples : Base de donnée de la *National Library of Medicine* interrogeable via *PubMed* [19], *Esp@ceNet*, la base de données européenne de brevets [20] ou *GlobalSpec* , base dédiée à l'information technique et d'ingénierie [21].

2. Les Bases de données internes à des sites web volumineux. Ces pages sont générées dynamiquement. Exemple : la base de connaissance des sites *Microsoft* [22].

3. Les publications, c'est-à-dire les Bases de données requêtables (via un moteur interne) donnant accès à des articles, des extraits d'ouvrages, des thèses, des livres blancs. Exemple : *FindArticles* [23].

4. Les sites de ventes en ligne, de ventes aux enchères. Exemple : Ebay, Fnac.com, Amazon.fr.

5. Les sites de petites annonces. Exemples : *de particulier à particulier*, *VerticalNet* (places de marché industrielles).

6. Les portails sectoriels. Portes d'entrées sur d'autres sites, ces sites agrègent plusieurs types d'information : articles, publications, liens, forums, annonces interrogeables via un moteur et une base de données. Exemple : *PlasticWay*, le portail dédié à la plasturgie [24].

7. Les bibliothèques en ligne. Bases de données issues essentiellement de bibliothèques universitaires ou nationales. Exemples : *Bibliothèques du Congrès US, BNF-Gallica*, les bases de bibliothèques recensées par l'*OCLC* [25] …

8. Les pages jaunes et blanches, les répertoires de personnes morales et physiques. Exemples : *Yellow Pages, ZoomInfo* (répertoire des dirigeants et salariés d'entreprises) [26]

9. Les Calculateurs, Simulateurs, Traducteurs. Ce ne sont pas des bases de données à proprement parler mais ces bases incluent de nombreuses tables de données pour calculer et afficher leurs résultats. Exemples : Traducteur *Systran Babelfish, Grand Dictionnaire Terminologique.*[27]

10. Bases de données d'emploi et de CV. Exemples : *Monster, APEC, Cadresonline*…

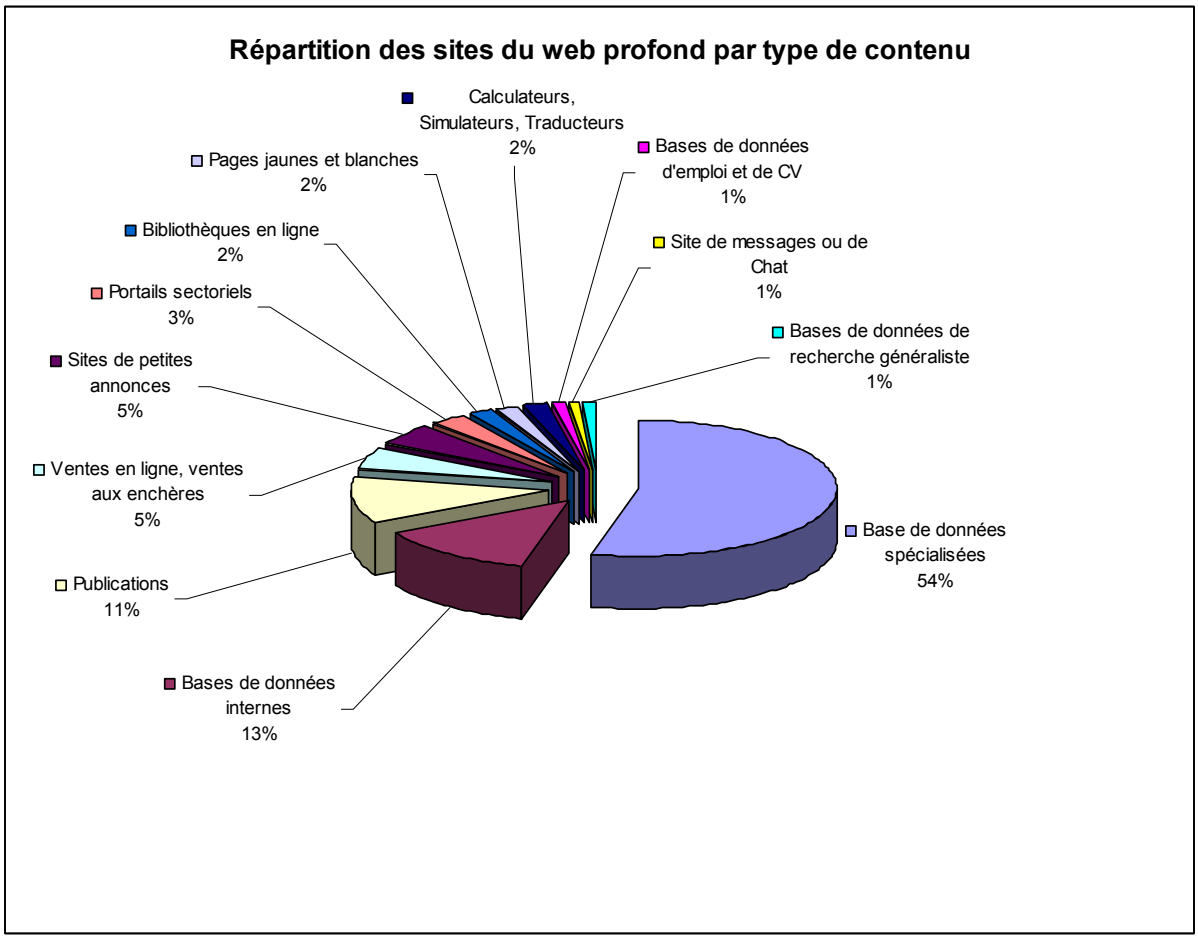
11. Site de messages ou de Chat

12. Bases de données de recherche généraliste. A la différence des bases de données spécialisées, elles recensent des thèmes éclectiques. Exemple : *Weborama*.

"Les bases de données à très forte notoriété - *Lexis Nexis, Dialog Datastar, Factiva*- ne constituent qu'à peine plus de 1% du web profond !"

Par ailleurs, l'étude révèle que 97,4 % de ces sites du web profond sont d'accès public, sans aucune restriction.

1,6% ont un contenu mixte : résultats disponibles gratuitement côtoyant des résultats nécessitant un enregistrement gratuit ou un abonnement payant.



Seulement 1,1 % des sites du web Invisible proposent la totalité de leur contenu par souscription ou abonnement payant. Ainsi, les bases de données à très forte notoriété telles que *Lexis Nexis, Dialog Datastar, Factiva, STN International, Questel*….ne constituent qu'à peine plus de 1% du Web Profond !

Couverture sectorielle du contenu des sites du web invisible

Les 17000 sites de l'étude révèlent une distribution relativement uniforme entre les différents domaines sectoriels. En fait, le web profond couvre tous les secteurs d'activités principaux :

Couverture du web profond	
Agriculture	2.7%
Arts	6.6%
Marchés	5.9%
Informatique/Web	6.9%
Education	4.3%
Emploi	4.1%
Ingénierie	3.1%
Gouvernement, services publics, collectivités territoriales	3.9%
Santé, Médecine, Biologie	5.5%
Sciences Humaines	13.5%
Droit et Politique	3.9%
Styles de vie	4.0%
Actualités, Media	12.2%
Personnes, Sociétés	4.9%
Divertissement, Sports	3.5%
Réferences	4.5%
Science, Maths	4.0%
Voyages	3.4%
Vente en ligne	3.2%

Répartition des sites du Web Profond par domaines sectoriels.© *Bright Planet* 2001

Qualité des sites du web invisible

La qualité est une notion évidemment subjective. Aussi, cette étude part d'un précepte simple : lorsque l'on obtient exactement les résultats que l'on désire via un site du web profond, on considère que la qualité dudit site est très bonne.

A l'inverse, si vous n'obtenez pas de résultats satisfaisants, la qualité est considérée comme nulle.

En terme de pertinence, la qualité du Web Profond est estimée **comme 3 fois supérieure à celle du web de surface.**

Pourquoi ? Essentiellement parce que la majeure partie des sites du web invisible sont des sites spécialisés, dédiés à une activité, une technologie, un métier et que leur contenu émane ou est validé par des professionnels, spécialistes et experts.

"Une qualité de contenu 3 fois supérieure à celle du web visible"

Les sites du web profond comportent donc des ressources plus volumineuses et de meilleure qualité que le web de surface. Mais comment accéder facilement à ces sites précieux ?

Comment identifier des sites du Web Invisible

On pourra trouver des sources de manière artisanales via des répertoires et moteurs spécialisés.

La plateforme *Digimind Evolution* vous permettra, elle, d'industrialiser d'une part, le processus de découverte et, d'autre part, d'automatiser la démarche de surveillance de sites du web invisible

Abordons tout d'abord les démarches "artisanales".

IDENTIFIER DES SOURCES VIA LES REPERTOIRES ET ANNUAIRES SPECIALISES

Dans la démarche de découverte de sites du Web Profond, certains sites du web visible peuvent vous aider. Il s'agit de moteurs dédiés au web invisible ainsi que des répertoires catégorisant des sites appartenant au web profond. Les présentations d'outils et répertoires ci-dessous ne prétendent pas être exhaustives. Elles entendent vous donner un aperçu de la richesse des ressources du web profond. Dans cette partie, nous vous présentons les outils de recherche spécialisés, les répertoires horizontaux, les répertoires spécialisés, les bases de données interrogeables, les bibliothèques en ligne et les portails sectoriels.

Les outils de recherche

- **Yahoo! Search Subscriptions**

Adresse : http://search.yahoo.com/subscriptions

Type : Moteur de recherche

Editeur : Yahoo!

Langue : anglais

Type d'accès : gratuit. Contenu complet payant

Avec Yahoo! Search Suscriptions, Yahoo! vous permet de rechercher parmi le contenu de publications de référence, d'études et de grands serveurs commerciaux. Une recherche vous apportera une liste de résultats comportant les titres et les extraits des articles ou études. Pour lire l'ensemble du contenu, il conviendra de vous abonner à la publication. Yahoo! Search Suscriptions donne notamment accès au contenu de : *Consumer Reports, FT.com, Forrester Research, IEEE publications, New England Journal of Medicine,TheStreet.com, Wall Street Journal, Factiva et LexisNexis.*

- **Gosh Me**

Adresse : http://www.goshme.com/

Type : Moteur de recherche

Editeur : GoshMe

Langue : anglais

Type d'accès : gratuit.

Selon le principe (pas toujours vrai), "A chaque type de requête, son moteur spécialisé", GoshMe vous permet de trouver pour vous les moteurs et répertoires qui correspondent le mieux à vos mot clés. Cet outil recherche les sources pertinentes parmi une sélection de plus de 1000 moteurs, répertoires et bases de données. Pour les moteurs, il convient de choisir parmi 17 grandes catégories : Animals & Farming, Arts & Social Sciences, Audio, Video & Images, Economics & Business, Environment & Society, Health, Science....(jusqu'à 5 en simultané). Ainsi pour la requête "brain" dans les catégories Sciences et Informatique, GoshMe propose 81 moteurs spécialisés, vous indique pour chacun le nombre de résultats pertinents et donne un lien vers chaque première page de résultats. Une approche "multimoteur" originale et très pratique...

- **Incywincy**

Adresse : http://www.incywincy.com/

Type : Moteur de Recherche

Editeur : Loop Improvements LLC

Langue : Anglais

Type d'accès : Gratuit

Incy Wincy crawle et indexe plus de 50 millions de pages, et possède une base de données de plusieurs centaines de milliers de moteurs de recherche. En crawlant à l'intérieur des sites, ce moteur construit un index de centaines de moteurs de recherche internes rencontrés. Exemple : Une requête sur la "CIA" permettra de détecter, sur les sites proposés par l'ODP, le moteur de l'United States Intelligence Community. Une recherche sur "Chimie" affichera, parmi les résultats les moteurs internes du Bottin de la Chimie et de l'Ecole Nationale Supérieure de Chimie de Rennes. On peut ensuite réutiliser ces moteurs par un simple clic de souris pour préciser la recherche. Une démarche originale.

• **FindArticles**

Adresse : http://www.findarticles.com/

Type : Base de données

Editeur : Find Articles

Langue : Anglais

Type d'accès : Gratuit et Payant

Ressources : Moteur de recherche interne, Indexation des magazines, Articles d'actualités, Classement d'articles par sujet.

Find Articles offre la possibilité de chercher parmi plus de 5.000.000 d'articles (remontant jusqu'à 1984), livres blancs, ou autres journaux et magazines. La recherche peut être effectuée par catégories (sujets ou journaux), ou grâce à un moteur de recherche. Ce dernier propose de rechercher par catégories ou par type d'accès au document (gratuit ou payant).

• **Google Scholar**

Adresse : http://scholar.google.com/

Type : Moteur de recherche

Editeur : Google

Langue : Anglais

Type d'accès : Gratuit

Google Scholar est un moteur de recherche spécialisé dans les documents académiques, les livres blancs, les articles scientifiques et méthodologiques. Il permet de rechercher par auteur, de trouver des citations, des livres et recense une grande partie des documents disponibles sur le web.

• **HighBeam**

Adresse : http://www.highbeam.com/library/index.asp

Type : Moteur de recherche

Editeur : HighBeam

Langue : Anglais

Type d'accès : Limité ou complet avec abonnement

Ressources : Recherche dans bibliothèque**,** Recherche dans le web**,** Recherche de références.

HighBeam permet de rechercher sur le net à travers trois onglets : Le premier, Library, permet de rechercher plus de 3.000.000 d'articles provenant de plus de 3.000 sources pouvant dater de plus de 20ans. La recherche permet de trouver des informations au travers du web, des news, des fils de discussion, des informations business ou encore des sites axés sur la recherche. Il est possible de customiser la recherche, de sauver les préférences ainsi que les précédentes recherches. L'onglet références permet quant à lui de chercher au travers de 43 encyclopédies, thésaurus, almanachs, dictionnaires pour atteindre les 295.000 références.

• **HighWire Press**

Adresse : http://www.highwire.org/

Type : Moteur de recherche

Editeur : Stanford University

Type d'accès : Gratuit (limité) ou complet (abonnement)

Langue : Anglais

Ressources : Biological Sciences, Physical Sciences, Medical Sciences, Social Sciences**,** Recherche par journal.

Moteur de recherche d'articles scientifiques, *HighWire* indexe plus de 15.000.000 d'articles au travers de 4.500 journaux issus de *PubMed,* incluant 900.000 articles gratuits. Il est possible de rechercher par auteur, mot clefs, année, volume ou encore nombre de pages parmi soit une sélection préalablement établie ou tout simplement parmi les 4.500 journaux de *HighWire*.

• **MagPortal**

Adresse : http://magportal.com/

Type : Répertoire

Editeur : Hot Neuron

Type d'accès : Gratuit, mais inscription nécessaire

Langue : Anglais

Ressources : Newsfeed (payant), Stockmarkets (payant), Newsletter, Personnalisation et capitalisation de la recherche.

Ce site permet de trouver des articles de magazine grâce à son moteur de recherche interne ou au travers de différentes catégories. Il présente les résultats les plus récents par rapport aux catégories choisies, et propose en même temps de trouver des articles plus anciens.

• **Scirus**

Adresse : http://www.scirus.com/srsapp/

Type : Moteur de recherche

Editeur : Scirus, groupe Reed Elsevier

Type d'accès : Gratuit

Langue : Anglais

Ce moteur de recherche est orienté sur les sources à contenu scientifique, sur le Web ou dans la presse. Il permet également d'identifier des sites d'universités, institutionnels, etc. Plusieurs fonctions sont disponibles telles la catégorisation, la conservation des requêtes et résultats ou encore la discussion par courrier électronique. *Scirus* est considéré comme le moteur de recherche scientifique de référence. Appartenant au groupe de presse Reed Elsevier, il indexe certains articles issus de revues comme *ScienceDirect*.

© Digimind

Les répertoires horizontaux

On parle d'approche horizontale, lorsque, par opposition à une approche verticale (par secteur), on recense des activités transversales (le droit, les moteurs de recherche).

Les répertoires de moteurs

• **Allsearchengines**.
Adresse: http://www.allsearchengines.com
Type : Répertoire
Type d'accès : Gratuit
Langue : Anglais

Ce répertoire recense, à travers 30 catégories, plus de 500 moteurs spécialisés par technologies ou secteurs (médecine, géographie, actualités, sciences, sport…)

• **FinderSeeker**
Adresse: http://www.finderseeker.com/
Type : Moteur
Type d'accès : Gratuit
Langue : Anglais

FinderSeeker recherche parmi des centaines de moteurs classés par thématique (automobile, géographie, sciences, religion…) ou pays.

• **Search Engine Colossus**
Adresse: http://www.searchenginecolossus.com/
Type : Répertoire
Editeur: Bryan Strome
Type d'accès : Gratuit
Langue : Anglais

Liste impressionnante de moteurs de recherche par pays. Une liste de 198 pays et 61 territoires est disponible. Au travers de chacune de ces catégories sont présentés les moteurs de recherche par pays.

• **Beaucoup**

Adresse : http://www.beaucoup.com/

Type : Répertoire

Editeur : T. Madden

Type d'accès : Gratuit

Langue : Anglais

Portail de moteurs de recherches et répertoires spécialisés, *Beaucoup* fédère plus de 2500 outils de recherche organisés en 15 grandes catégories (Geographical, Health, Media, Business, Society...).

Répertoires de Presse

• **All Newspapers**

Adresse : http://www.allnewspapers.com/

Type : Répertoire de Journaux

Editeur : Allnewspapers

Type d'accès : Gratuit

Langue : Anglais

Répertoire de ressources sur les journaux du monde entier. Le classement est effectué par continent, puis par pays et enfin par région. Différents types de médias sont accessibles comme la presse écrite mais aussi les sites de télévision ou encore de radio

• **Online Newspapers**

Adresse : http://www.onlinenewspapers.com/.

Type : Répertoire de Journaux

Editeur : Web Wombat

Type d'accès : Gratuit

Langue : Anglais

Ce portail permet d'accéder à des milliers de journaux nationaux et régionaux. La recherche peut s'effectuer par grandes zones géographiques (USA, Asie Pacifique, Asie du Sud Est, Europe, …) ou pays par pays (de l'Afghanistan au Zimbabwe). Très clair.

• **Giga Presse**

Adresse : http://www.presse-on-line.com/

Type : Répertoire de Journaux

Editeur : Giga Presse

Type d'accès : Gratuit

Langue : Français

Ressources : Répertoire de journaux et magazines, Recherche d'articles**,** Revue du Net**,** Actualité du jour.

Ce portail est en fait un répertoire dédié aux journaux et magazines. Ceux-ci sont classés par catégories (art & culture, divertissement et loisirs, actualités, enseignement, …). Ces journaux sont catégorisés en fonction de la pertinence attribuée par l'éditeur.

Les répertoires spécialisés

Contrairement aux répertoires généralistes tels que *Yahoo! Directory*, les répertoires spécialisés sélectionnent pour leur qualité certaines ressources de référence. L'approche n'est donc pas quantitative mais focalisée sur la pertinence et la spécialisation des sites sans souci d'exhaustivité.

• **About.com**

Adresse : http://www.about.com

Type : Répertoire

Editeur : About Inc, groupe New York Times

Type d'accès : Gratuit

Langue : Anglais

Ressources : Chaînes thématiques**,** Navigation par sujet**,** Newsletter**.**

About.com est considéré comme le répertoire de référence. Ce répertoire est basé sur 23 chaînes thématiques qui fournissent des informations originales sur plus de 475 sujets différents. Ces sujets sont renseignés à l'aide de "guides" spécialisés, qui proviennent du monde entier, et qui ont une approche d'expert. De nombreux articles y sont ainsi rédigés, et donnent à *About.com* une forte légitimité sur le web.

• **Resources Discovery Network**

Adresse : http://www.rdn.ac.uk

Type : Répertoire

Editeur : JISC

Type d'accès : Gratuit

Langue : Anglais

Ressources : Répertoire, Publication, Etude de cas, Formation.

RDN est une porte d'entrée britannique (gateway) d'un réseau de portails consacrées à une douzaine de grands thèmes représentant plus de 100 000 ressources : Affaires (Business), Informatique, Ingénierie, Mathématiques, Sciences Physiques, Sciences Sociales, Sciences Humaines, Droit et à des Sujets de référence (bibliothèques, périodiques) et dans un second temps l'Education, Géographie et Sports. Le RDN est organisé comme une coopérative avec une organisation

centrale (RDN) et des fournisseurs indépendants appelés les hubs ("pivots répartisseurs") comprenant notamment :

BIOME : Health and Life Sciences

EEVL : Engineering, Mathematics and Computing

Humbul : Humanities

PSIgate : Physical Sciences

SOSIG : Social Sciences, Business and Law

On peut accéder à chacun de ces hubs de manière indépendante ou en utilisant l'interface de recherche du RDN.

Autres fonctions :

RDN's "Virtual Training Suite" : tutoriaux pour aider les étudiants et chercheurs

RDN "Behind the Headlines" : liens vers des ressources très spécialisées "délaissées" par les grands médias traditionnels.

- **Infomine**

Adresse : http://infomine.ucr.edu

Type : Moteur de recherche

Editeur : Library of the University of California

Type d'accès: Gratuit

Langue: Anglais

Infomine se définit comme une bibliothèque virtuelle qui possède une importante collection de liens annotés (plus de 100.000). Cette collection comprend des bases de données, des journaux électroniques, des guides Internet pour toutes les disciplines ou encore des abstracts de conférences.

Le moteur de recherche permet une recherche fine et pointue de documents grâce à un module de recherche avancé qui donne accès à plusieurs champs à remplir aidant la recherche.

- **Bubl Link**

Adresse : http://bubl.ac.uk/

Type : Répertoire

Editeur : StrathClyde University

Type d'accès : Gratuit

Langue : Anglais

Ressources : Répertoire classé selon *Dewey*, Moteur de recherche interne.

Bubl Link est un répertoire de ressources Internet (plus de 12000) à vocation scolaires et universitaires : sites web de référence, mailing lists, newsgroups, bases de données dédiés aux Arts, Sciences Humaines, Littérature, Sociale, Mathématiques, Santé, Ingénierie, Physique... Chaque grand domaine est subdivisé en centaines de sous-catégories.

• **Invisible Web.net**

Adresse : http://www.invisible-web.net/

Type : Répertoire

Editeur : Gary Price, Chris Sherman

Type d'accès : Gratuit

Langue : Anglais

Ressources : Répertoire**,** Livre disponible (résumé en pdf)

Chris Sherman (SearchDay) et Gary Price (ResourceShelf, DirectSearch), parmi les meilleurs experts américains des moteurs de recherche, (*www.searchenginewatch.com*) ont réalisé ce répertoire qui propose plus de 1000 sources d'accès au web invisible. C'est en fait la version on-line et la mise à jour du répertoire "papier" de la seconde partie de leur ouvrage "*The Invisible Web : Finding Hidden Internet Resources Search Engines Can't See*"**.**

• **Librarians Index to the Internet**

Adresse : http://lii.org/

Type: Répertoire

Editeur: IMLS

Type d'accès: Gratuit

Langue : Anglais

Ressources : Répertoire**,** Moteur de recherche**,** Syndication**,** Nouveaux sites.

Répertoire de près de 16 000 ressources internet sélectionnées par les documentalistes. Le *LII* a été fondé en 1990 et a migré en 1993 vers le serveur de la bibliothèque publique de Berkeley (CA). Le moteur de recherche interne a été développé à partir de 1996. Ce moteur permet de rechercher par sujet, titre, contenu ou URL. La recherche avancée permet d'utiliser les booléens AND, OR et NOT et restreindre le champ des recherches aux bases de données, répertoires, "best of" ou ressources spécifiques. La rubrique "Browse All Subjects" permet

d'afficher tous les thèmes classés par ordre alphabétique. On peut également effectuer la recherche via les thématiques (de Arts à Women…).

• **Complete Planet**

Adresse : http://www.completeplanet.com

Type : Répertoire

Editeur : Bright Planet

Type d'accès : Gratuit

Langue : Anglais

Ressources : Moteur de recherche, Répertoire, Livres blancs.

Portail de recherche par mots-clés ou annuaire thématique sur plus de 90 000 bases de données ou moteurs de recherche spécialisés (de Agriculture à Weather). Interface et graphisme très agréable, possibilité de sauvegarder ses recherches, moteur de recherche puissant.

• **Direct Search**

Adresse : http://www.freepint.com/gary/direct.htm

Type : Répertoire

Editeur : Gary Price

Type d'accès : Gratuit

Langue : Anglais

Gary Price a compilé des centaines de liens vers des ressources du web invisible. Ce site de la George Washington University propose un accès par recherche directe ou par catégories thématiques (géographie, histoire, droit, cinéma...). Le bas de la page indique les récents ajouts. Ce n'est donc pas un répertoire au sens strict mais une liste très utile.

• **Internet Public Library**

Adresse : http://www.ipl.org/

Type : Répertoire

Editeur : University of Michigan
Type d'accès : Gratuit
Langue : Anglais
Ressources : Répertoire, Nouveautés, Blog

Plus de 60 000 ressources Internet organisées par les bibliothécaires et étudiants de l'université. Répertoire d'ouvrages de références, journaux, critiques littéraires, magazines, journaux et newsletters on-line, associations (essentiellement à but non lucratif). Recherche par moteur ou thématiques.

• **Scout Report**
Adresse : http://scout.wisc.edu/Archives/
Type : Moteur de recherche et Répertoire
Editeur : ISP
Type d'accès : Gratuit + abonnement
Langue : Anglais

Les archives de *Scout Report* proposent plus de 17 000 sites web et mailings lists sélectionnés pour leur contenu éducatif. Moteur de recherche par sujet, titre, résumé ou URL. Recherche avancée par booléens. *The Scout Report* est publié tous les vendredi depuis 1994 et propose une sélection de ressources Internet dans tous les sujets à valeur éducative. 3 lettres sont spécialisées en Sciences Physiques et Ingénierie, Commerce et Economie et Sciences Sociales.

• **All Academic**
Adresse : http://www.allacademic.com/
Type : Répertoire
Editeur : All Academic INC
Type d'accès: Payant
Langue: Anglais
Ressources : Moteur, Archives, Journal.

Répertoire de ressources scolaires et universitaires : plus de 200 publications, journaux et magazines sur plus de 50 domaines spécialisés. Recherche directe par mots-clés (auteur, titre, sujet). Liste de plus de 300 publications par disciplines :

Agriculture, Anthropologie, Géographie, Art, Architecture, Astronomie, Biologie, Commerce, Chimie, Communications, Informatique, Ethnologie, Sciences Economiques, Enseignement, Ingénierie, Anglais, Environnement et Ecologie, Cinéma, Folklore, Géologie, Géophysique, Gouvernement, Santé, Sciences Humaines, Histoire, Journalisme, Linguistique, Droit, Littérature, Mathématiques, Histoire médiévale, Musique, Philosophie, Physique, Sciences Politiques, Population et Démographie, Psychologie, Services Publiques, Religion, Sciences (général), Sciences Sociales(général), Travail, Sociologie, Théâtre, Urbanisme, ...

• **OAIster**

Adresse : http://oaister.umdl.umich.edu/o/oaister/

Type : Répertoire

Editeur : University of Michigan

Type d'accès : Gratuit

Langue : Anglais

Répertoire de ressources numériques comprenant plus de 5.384.000 documents provenant de 475 institutions (journaux, magazines spécialisés, etc.). Le moteur de recherche interne est particulièrement bien développé, et les résultats se présentent d'une manière à la fois très complète et intuitive. Il est possible de classer les résultats par titre, auteur ou encore dates, et les documents font l'objet d'une explication approfondie. Les champs remplis sont le titre, auteur, publieur, date, type de la ressource, son format, un résumé, le sujet, l'adresse et enfin l'institution qui la publie. De plus, sur la gauche de l'écran, les résultats sont présentés en fonction de la source (institution).

• **Allinfo**

Adresse : http://www.allinfo.com

Type : Répertoire

Editeur : Allinfo Inc

Type d'accès : Gratuit

Langue : Anglais

Ressources : Répertoire**,** Moteur de recherche**,** Visualisation graphique des catégories

Allinfo est un répertoire qui se divise en 4 grandes catégories : Science, Humanities, Social Science, Recreation. Ces grandes catégories sont elle aussi subdivisées en sous catégories. Il est ainsi possible de trouver une multitude de ressources au travers de ce site qui indexe une multitude de liens.
De plus, la présentation des catégories en fonction d'un mot clef est tout à fait intuitive grâce à une interface graphique qui permet de visualiser les catégories sous forme de carte.

Les bases de données interrogeables

A côté des bases de données payantes que sont *Dialog*, *Factiva* ou *Lexis Nexis*, il existe sur le web de nombreuses bases de données interrogeables à accès gratuit. Ces bases proposent de rechercher parmi leur contenu grâce à des interfaces avancées de recherche.

• **URFIST Lyon**

Adresse : http://dadi.enssib.fr/

Type : Répertoire de Bases de données

Editeur : URFIST

Type d'accès : Gratuit

Langue : Français

L'URFIST (Unité Régionale de Formation et de Promotion pour l'Information Scientifique et Technique) de l'Université Lyon I propose une sélection de près de 800 bases de données gratuites. Parmi les catégories accessibles, l'agriculture bien sûr mais également les brevets, les marques, la chimie, l'environnement, l'économie, la génétique, l'histoire, l'informatique, la linguistique, les mathématiques, la médecine, les sciences et puis le cinéma, l'art, la photographie...

• **URFIST de Nice**

Adresse : http://www.unice.fr/urfist/URFIST-DEH/pages/database.html

Type : Liste de liens

Editeur : URFIST

Type d'accès : Gratuit

Langue : Français

L'URFIST de l'Université de Nice-Sophia Antipolis propose elle aussi une sélection de bases de données gratuites.

• **The Internet Archive**

Adresse : http://www.archive.org/

Type : Base de données

Editeur : IA
Type d'accès : Gratuit
Langue : Anglais

The Internet Archive est une bibliothèque numérique destinée à conserver tous les documents issus de l'Internet pour les préserver d'une disparition complète. The *IA* fournit des documents créés à partir de 1996 (40 milliards de pages web mais aussi Usenet, films et l'ancêtre Arpanet). The Internet Wayback Machine (développée notamment avec Alexa Internet) permet à l'utilisateur de trouver des sites web archivés en tapant simplement son URL et la date désirée. Etonnant et très utile compte tenu de la disparition fréquente de sites internet. Accessible au public depuis le 24 octobre 2001.

Les bibliothèques en ligne

• **LibDex**

Adresse : http://www.libdex.com

Type : Répertoire

Editeur : Peter Scott University of Saskatchewan (Canada). Bisca International Investments Ltd

Type d'accès : Gratuit

Langue : Anglais

Répertoire de plus de 18000 bibliothèques publiques mais aussi privées à travers le monde (133 pays de l'Albanie au Zimbabwe) proposant un index de leur contenu en ligne. La recherche peut s'effectuer par pays mais également par Open Access Catalogs (OPAC) c'est-à-dire les catalogues informatisés signalant les ouvrages et les périodiques présents dans la bibliothèque. Egalement des liens vers les sites de e-commerce proposant des ouvrages en ligne et une liste des magazines, journaux et newsletters consacrés aux bibliothèques. Très complet.

• **Les catalogues de la Bibliothèque Nationale de France (BnF)**

Adresse : http://www.bnf.fr/pages/catalogues.htm

Type : Répertoire

Editeur : BNF

Type d'accès : Gratuit

Langue : Français

Cette page recense les catalogues de la *BnF*. Ces catalogues décrivent les documents et objets conservés à la BNF (documents imprimés, documents audiovisuels, cartes et plans, monnaies et médailles, manuscrits). Certains sont numérisés et/ou microfilmés :

- BN-OPALE PLUS, le catalogue des collections patrimoniales
- Catalogue des imprimés en libre-accès
- BN-OPALINE : le catalogue en ligne des collections spécialisées
- Catalogue des documents numérisés

• **OCLC**

Adresse : http://www.oclc.org

Type : Répertoire

Editeur : OCLC Online Computer Library Center

Type d'accès : Gratuit

Langue : Anglais

Réseau informatisé de bibliothèques fondé en 1967 par les présidents des collèges et universités de l'état d'Ohio (Ohio College Library Center).Près de 100 bases de données utilisées par 41 000 bibliothèques dans 82 pays (en France dans plus de 90 bibliothèques universitaires ou publiques).

- Les bases :

ECO (Electronic Collection Online) : accès aux revues électroniques

Worldcat : 45 millions de notices bibliographiques depuis 30 ans en 400 langues

NetFirst : base de sites web indexés et mis à jour

Fastdoc, ContentsFirst : bases de sommaires de périodiques

Sitesearch : Bibliothèque virtuelle.

- Outils de recherche :

Firstsearch : accès à des bases de données internationales

Firstsearch eco : accès à plus de 2500 titres de revues électroniques.

Les portails sectoriels

Leur approche est verticale : Ce sont des portails spécialisés dans un secteur d'activité, une technologie. On peut aussi parler de "Vortail" (contraction de "vertical" et "portail"). Ils sont nombreux. Ainsi pour le secteur de la chimie, il existe (entre autres...) *France-Chimie, Chemindustry, Chem.com, Chemscope, Chemweb, Chemicalonline* [28] ...auxquels il faut ajouter tous les portails concernant les secteurs dérivés de l'industrie chimique (plasturgie, pharmacie, cosmétique, pétrole, agrochimie,...)

• **Les 1000 meilleurs portails d'affaires sectoriels**

Adresse : http://www.objectifgrandesecoles.com/pro/secteurs/index.htm

Type : Répertoire

Editeur : Objectif grandes écoles

Type d'accès : Gratuit

Langue : Français

Le site *Objectif Grandes Ecoles* propose une rubrique consacrée aux portails sectoriels. Sont ainsi abordés les secteurs : Agriculture et Agro-alimentaire, Bâtiment et Construction, Commerce et distribution, Energie et ressources naturelles, Biens de consommation, Informatique et Télécommunications, Industries lourdes, Médias et journalisme, Professions artistiques, loisirs et tourisme, Santé et Environnement, Secteur public, enseignement et recherche, Services financiers et consulting, Transport et logistique.

Mais, le point faible de ces moteurs, annuaires et répertoires décrits ci-dessus est le fait de proposer une recherche sur des catégories déjà présélectionnées, et donc ne correspondant pas obligatoirement à votre domaine d'activité spécifique. D'autre part, la consultation de toutes ces ressources est chronophage.

Aussi, il peut être plus rentable et efficace de chercher directement par vous-même, via des requêtes précises, les ressources dédiées à votre problématique…

RECHERCHER DES RESSOURCES SPECIALISEES VIA VOTRE REQUETE

Pour entamer une recherche plus précise sur une technologie, un secteur, pour trouver le portail vertical qui concerne votre domaine, il peut être nécessaire d'effectuer une recherche avec vos propres mots clés.

Si les moteurs généralistes comme *Google* ou *Yahoo!* n'indexent pas le contenu de très nombreuses bases du web profond, ils peuvent néanmoins vous aider à trouver des "métaressources", à savoir des listes de liens, des répertoires, des portails ou des bases de données spécialisées dans votre métier.

Comment construire une requête efficace pour identifier ces métaressources ?

Choisir ses mots clés

Liste et choix des mots clés de votre secteur

Il convient tout d'abord de lister les mots clés propres à votre secteur ou problématique;

A savoir : les concepts, verbes, synonymes, concepts liés, acronymes….

Si votre métier est très spécifique, risquant ainsi de générer peu de réponses, pensez à imaginer les termes qui permettent de l'élargir. Ainsi, par exemple, la Biotechnologie découle de la biologie et decertains secteurs high-tech, l'Hydrodynamique est une branche de la Mécanique des fluides.

Liste des supports et des types de contenu attendus

Listez ensuite les mots qui vont décrire les formats de ressources attendues, et ce, en plusieurs langues :

- liens / links, ressources/sources, bases de données/database, white paper/ livre blanc, technical paper, bookmarks, directories, etc…

Construire les équations de recherche

Dans cette démarche, il ne s'agit plus d'effectuer une recherche dans l'optique de trouver immédiatement l'information elle-même mais dans le but de déceler des métaressources spécialisées (susceptibles de proposer des sites web pertinents).

Pour cela, il convient de multiplier les associations de vos mots clés "secteurs" à vos mots clés "supports/contenu" :

Exemple : *industrie aéronautique* + *ressources / sources / liens / annuaire, répertoire / directories…*

Ce type de requête est à effectuer sur des moteurs à large index tels que *Google, Yahoo! Search, Gigablast, MSN Search* ou *Exalead* [1 et 6].

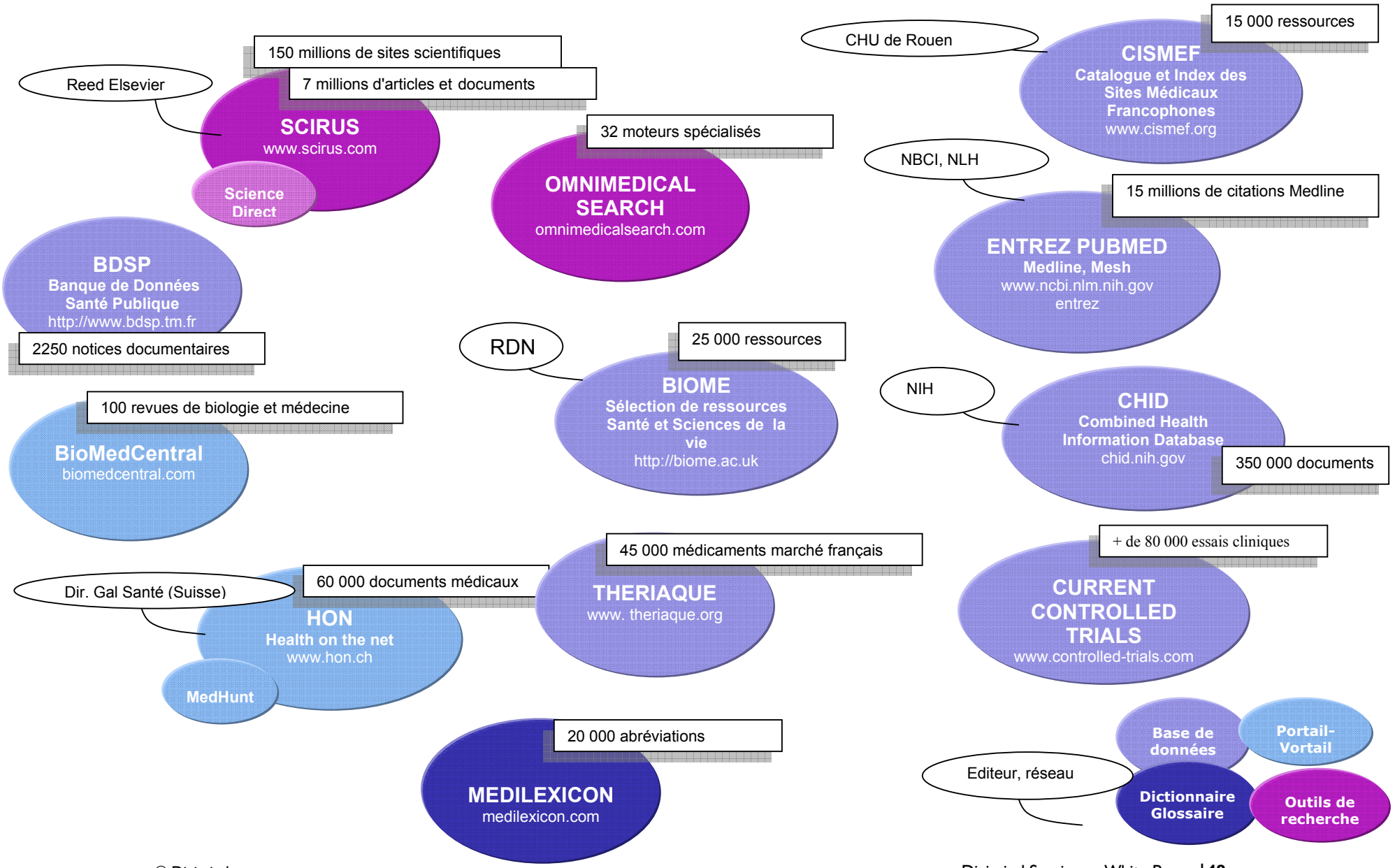
Mais cette approche, bien que souvent riche en résultats, reste toutefois très longue et fastidieuse. Il faut en effet multiplier les équations de recherches (faire varier les associations de mots, les langues, les singuliers/pluriels…) sur plusieurs moteurs.

Compte tenu de ce processus chronophage, Digimind à développé de nouveaux outils pour industrialiser et simplifier les processus de recherche et de surveillance sur le Web Invisible.

EXEMPLE DE RESSOURCES DU WEB INVISIBLE

La page suivante présente une cartographie d'exemples de ressources du web invisible pour le secteur Santé (Médecine – Pharmacie – Biologie). Il s'agit d'un échantillon de sites non indexés ou partiellement indexés par les moteurs conventionnels.

DIGIMIND



© Digimind

Digimind et la veille sur le Web Invisible

LES SPECIFICITES D'UNE VEILLE AUTOMATISEE SUR LE WEB INVISIBLE

Le web profond présente un certain nombre de caractéristiques qui nécessitent des technologies de surveillance adaptées, si l'on souhaite pouvoir effectuer une veille efficace.

En effet, un veilleur qui doit consulter régulièrement une vingtaine de moteurs sur plusieurs dizaines de thématiques risque d'y passer ses journées.

D'où la nécessité d'utiliser des outils de "méta-recherche" et de surveillance de ces moteurs de recherche pour automatiser autant que possible le travail répétitif d'identification et de rapatriement de nouvelles informations pertinentes apparues sur le web invisible.

Afin d'apporter un maximum de productivité dans le travail du veilleur, ces outils doivent surmonter les obstacles suivants :

- L'interfaçage avec des moteurs internes de sites de tout type ;

- La nécessité d'extraire uniquement – mais intégralement - les résultats des moteurs ;

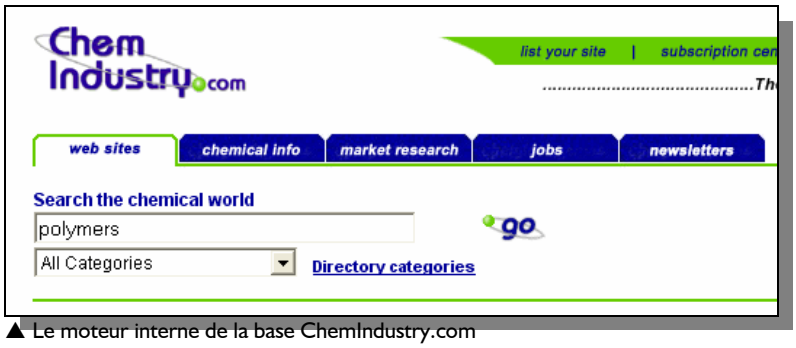
- La possibilité d'explorer l'intégralité de l'information "derrière les résultats" ;

- La nécessité de filtrer les résultats redondants ;

- La nécessité de traiter de gros volume de données.

L'interfaçage avec des moteurs internes de sites de tout type

La surveillance du web profond nécessite d'interroger les moteurs internes de sites: moteur de portail, moteur de bases de données. Ces moteurs utilisent des modes de requêtage très variés, depuis le type de requête (GET, POST, mixte), le format des formulaires d'interrogation (HTML, Javascript), la méthode de soumission de la requête (formulaire ou redirection Javascript), ou encore la syntaxe de la requête. L'interfaçage avec ces différents moteurs est un processus qui peut s'avérer particulièrement complexe afin de mémoriser puis automatiser les requêtes posées.



▲ Le moteur interne de la base ChemIndustry.com

Les métamoteurs :

Les métamoteurs du web « visible » sont légions : *Dogpile, Mamma.com, Search.com*. La plupart recherchent dans les moteurs « ouverts » les plus connus, tels que Google, Yahoo, AskJeeves, MSN Search ou LookSmart.

Moins nombreux sont les métamoteurs qui permettent de rechercher dans les bases et moteurs du web invisible: *Profusion, Turbo10* ou *Copernic Agent*.

En conséquence, la plupart des métamoteurs du web invisible (voir ci-contre) ne recherchent que dans leur propre sélection de moteurs pré paramétrés (tel Copernic Agent ou Turbo 10). Copernic annonce ainsi un chiffre de 1000 moteurs regroupés en 120 catégories…tandis que Turbo10 permet de rechercher dans des groupes de 10 moteurs, choisis parmi 795 répertoriés. Un choix qui reste néanmoins limité en regard des 200.000 sites du web invisible répertoriés dès 2001 ! Utile pour une veille généraliste, ces outils montreront vite leurs limites pour le veilleur professionnel ou l'expert de l'entreprise.

De leur côté, les logiciels de surveillance de sites web permettant d'accéder au Web Invisible (tels KB Crawl ou WebSite Watcher) contournent la difficulté de création d'un "connecteur générique" à chaque moteur en mettant en surveillance la page de résultats de chaque requête – ce qui nécessite le paramétrage manuel de chaque requête sur chaque moteur, travail qui peut s'avérer extrêmement fastidieux pour notre veilleur interrogeant une vingtaine de moteurs sur plusieurs dizaines de thématiques (20 moteurs x 90 thématiques = 1800 requêtes à paramétrer…et administrer).

© Digimind

L'approche de Digimind : configurer des connecteurs en 3 clics

Pour permettre de mettre rapidement en surveillance toute nouvelle requête sur sa propre sélection de moteurs, Digimind a développé des technologies exclusives de configuration de connecteurs en trois clics.

La première reconnaît automatiquement les formulaires html et javascript, tandis que la deuxième permet l'apprentissage sémantique de la requête http. Résultat : le veilleur a juste besoin de simuler une recherche sur ce moteur pour que le connecteur soit activé…et le moteur immédiatement disponible pour des recherches en temps réel ou les surveillances en cours.

La nécessité d'extraire uniquement – mais intégralement - les résultats des moteurs

Tout comme leur interface de recherche, les pages de résultats des différents moteurs varient grandement d'un moteur à l'autre en terme d'affichage:

- éléments constituant les résultats (titre, date, résumé pour des actualités, nom de brevet, inventeur, numéro de brevet pour des bases de brevets, etc…) ;
- position des résultats dans la page (voire listing des résultats dans une nouvelle page) ;
- informations non pertinentes apparaissant autour des résultats (rubriques, publicités, aide, contacts,…)

Les logiciels classiques de surveillance du web :

Ces logiciels, parmi lesquels *Copernic Tracker, KB Crawl* ou encore *WebsiteWatcher*, identifient les modifications apparaissant sur les sites et pages web en comparant la nouvelle version d'une page avec l'ancienne et en surlignant tous les nouveaux éléments.

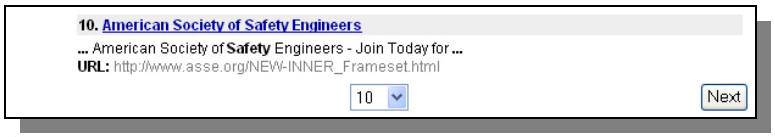
Certains d'entre eux sont capables de mettre en surveillance une requête sur un moteur de recherche, mais restent dans l'impossibilité de distinguer les résultats du reste des contenus de la page de résultats, ou de suivre spécifiquement les pages de résultats suivantes parmi tous les liens présents sur la page.

▲ *La page de résultats d'Esp@ceNet : la base du Bureau Européen des Brevets*

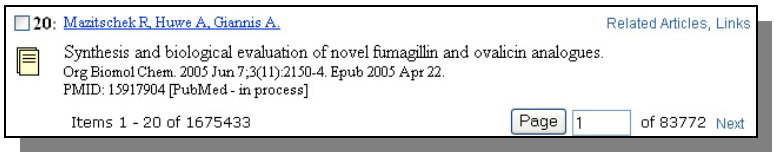
Là encore, les métamoteurs les plus courants auront préparamétré leur propre sélection de moteurs en créant un fichier descriptif pour chaque page de résultats. Ce "masque" indique clairement au "robot" du métamoteur à quelles balises HTML commence et se termine le listing de résultats, et entre quelles balises HTML se situent les éléments constitutifs de chaque résultat. Ce fichier descriptif étant fastidieux à créer, on comprend là encore pourquoi ces métamoteurs ne permettent pas à leurs utilisateurs d'ajouter leurs propres moteurs.

Le problème corollaire de la configuration "manuelle" de chaque page de résultats est qu'en cas de modification de la structure HTML de la page de résultat par l'éditeur du site, le fichier descriptif cesse de fonctionner…et doit être mis à jour.

L'autre difficulté à surmonter est de suivre, le cas échéant, les multiples pages de résultats (« *page 1, page 2, page 3, pages suivantes*… »). Là encore, les métamoteurs nécessitent la configuration manuelle d'un fichier décrivant le mode de navigation entre les pages de résultats, navigation qui peut s'avérer facilement identifiable lorsque les numéros de pages ou de résultats apparaissent en clair dans l'adresse de la page de résultats (URL), mais beaucoup moins évidente dans les autres cas.



▲ *Afficher les 10 résultats suivants de Chemindustry.com*



▲ *Afficher les pages suivantes sur PubMed.com*

Gérée moteur par moteur à l'aide de fichiers descriptifs par les métamoteurs, l'extraction des résultats n'est absolument pas gérée par les logiciels classiques de surveillance de pages. Ceux-ci sont en effet incapables de distinguer la zone de résultats du reste des informations de la page – générant en conséquence beaucoup de "bruit". Qui plus est, ils crawleront les autres pages de résultats comme n'importe quel autre lien sur la première page de résultat, rapatriant pour chaque nouvelle page de résultats crawlée de multiples pages non pertinentes.

L'approche de Digimind : la reconnaissance et l'extraction automatique

Là encore, Digimind a développé des technologies exclusives de reconnaissance et d'extraction automatiques des résultats de moteurs de recherche. Basée sur des algorithmes avancés de reconnaissance de forme, cette technologie présente le grand avantage de s'adapter automatiquement à tout moteur de recherche – et même à toute modification d'un moteur de recherche.

Notre veilleur peut ainsi consulter et surveiller à l'aide de la plateforme Digimind Evolution tous les résultats, et uniquement les résultats, des moteurs interrogés – sans aucune intervention de sa part.

Pouvoir explorer l'information "derrière les résultats"

Comme indiqué précédemment, les résultats des moteurs de recherche sont constitués de quelques éléments les plus représentatifs de l'information complète à laquelle ils mènent : titre, date, auteur, résumé dans la plupart des cas…ou encore notice synthétique dans le cas de bases documentaires structurées.

Dans le cadre d'une surveillance, ce résumé compte bien sûr, mais moins que l'intégralité de l'information vers laquelle il renvoie. En effet, les mots clés de votre surveillance peuvent tout à fait se situer au fin fond d'un article scientifique de 90 pages, sans que la notice seule puisse vous le révéler – voire sans que ce document ait été indexé dans son intégralité (voir page 8 les limites d'indexation des moteurs).

Il est donc indispensable dans le cadre de l'automatisation d'une veille du Web Invisible que les outils utilisés soient capables de lire l'intégralité des informations "derrière les résultats".

Cela suppose également que ces outils soient capables de lire des documents en d'autres formats que le html, car ceux-ci sont très fréquemment publiés directement en Pdf, Word, Excel, Powerpoint ou encore postscript.

Les principaux métamoteurs du web invisible ont une limitation majeure par rapport à ces spécificités de la surveillance du web invisible : ils se contentent la plupart du temps de rapatrier les résultats des moteurs de recherche et bases interrogées, sans aller crawler et indexer les informations derrière les liens des résultats. Du coup, ils sont limités par l'indexation effectuée par les moteurs de recherche, elle-même généralement non exhaustive (500 Ko par document pour Google et Yahoo ! par exemple).

Quant aux logiciels de surveillance du web, ils seront confrontés aux mêmes limitations que précédemment : étant incapables de distinguer ni les résultats, ni les pages suivantes des résultats, des autres liens de la page, ils crawleront sans distinction tous les liens, rapatriant pour chaque document pertinent, de multiples informations non pertinentes.

L'approche de Digimind : explorer, extraire et unifier automatiquement l'intégralité des informations derrière les résultats

Là encore, l'approche de Digimind se distingue par des technologies uniques d'exploration et d'homogénéisation de tout type de documents présents derrière les liens des résultats. Une fois la zone des résultats et le mode de navigation entre les pages identifiés par le 1er algorithme, et les résultats décomposés en leurs éléments constitutifs par un 2ème algorithme (voir plus haut) - dont les liens vers les documents cibles des résultats - chacun de ces derniers est alors exploré par un puissant robot. Celui-ci en effet :

- gère les formats bureautiques les plus répandus (pdf, doc, rtf, ppt, xl, postscript, en plus du html, xml, et txt)

- gère tous les types d'encoding de documents, dont non-occidentaux (japonais, chinois, russe, arabe,…)

- explore l'intégralité du document cible, sans limitation de taille

- génère à la volée un résumé des parties les plus pertinentes par rapport à la requête mise en surveillance

Résultat : une surveillance profonde et exhaustive du web invisible, quelles que soient les langues ou les formats des documents cibles.

La nécessité de filtrer les résultats redondants

L'interfaçage de plusieurs moteurs internes de sites du Web Invisible peut vous amener des informations redondantes, en double voire en triple. En effet, certains documents, publications et analyses peuvent être présentes sur plusieurs bases ou portails différents.

Gérée par la plupart des métamoteurs en regroupant les résultats identiques – mais rarement les informations similaires, puisqu'ils ne les re-explorent pas (voir plus haut) – la redondance n'est pas du tout prise en compte par les logiciels de surveillance du web. Dans leur cas, chaque requête effectuée sur différents sites du web invisible rapportera son lot d'alertes redondantes, sans possibilité de regroupage de ses informations autre que manuellement.

L'approche de Digimind : regrouper automatiquement les informations similaires

Re-explorant les documents cibles des résultats des recherches, le robot de Digimind est capable de regrouper entre eux tous les documents identiques.

La nécessité de traiter de gros volume de données

Nous l'avons vu, le Web Profond est constitué de sites en moyenne plus volumineux que les sites du web visible (60% des sites les plus vastes du Web Profond

représentent à eux seuls un volume qui excède de 40 fois le Web de Surface). De plus, le volume du web profond à un taux de croissance très élevé (multiplication par 9 chaque année). Enfin, nous avons compris l'importance de pouvoir explorer l'intégralité des documents issus des résultats de recherches, afin de ne pas rater une information clé située au fin fond d'un fichier PDF de plusieurs mégaoctets.

Dès lors, la puissance nécessaire à la surveillance du web invisible devient hors de portée de logiciels monopostes installés sur son PC de bureau. Copernic limite ainsi à 700 le nombre de résultats par moteur interrogé - et il ne s'agit d'afficher que les résultats, sans explorer les documents indexés. Même les métamoteurs grands publics tournant sur les index de Google, Yahoo ! ou MSN Search limitent le nombre de requêtes simultanées ou le nombre de moteurs interrogés simultanément.

L'approche de Digimind : puissance, ciblage et profondeur

Les logiciels de Digimind sont hébergés sur des serveurs puissants et leur conception en J2EE autorise une montée en charge infinie en reliant les serveurs en clusters.

D'autre part, permettant de se constituer ses propres bouquets de sources du web invisible, la technologie Digimind évite la recherche simultanée dans des milliers de moteurs non pertinents. A une veille horizontale large et généraliste, l'approche préconisée est celle d'une veille ciblée (sur les sources réellement pertinentes pour le veilleur concerné) et profonde (en explorant l'intégralité des informations).

La puissance est donc mise au service de l'exploration intégrale d'une sélection pertinente de sources.

COMPARATIF DES PRINCIPAUX OUTILS DE RECHERCHE / SURVEILLANCE DU WEB INVISIBLE*

Solutions	Digimind Evolution	Logiciel classique de surveillance du web	Métamoteur "web invisible"
Fonctionnalités clés			
Possibilité de créer en 3 clics des connecteurs génériques vers les moteurs de son choix	**OUI**	NON	NON
Possibilité de mise en surveillance de sa requête sur un bouquet de moteurs	**OUI** (illimité)	NON (requête à resaisir moteur par moteur)	OUI
Extraction de multiples pages de résultats	**OUI** (illimité)	NON (crawling indifférencié)	OUI
Extraction automatique des résultats	**OUI**	NON	OUI
Dédoublonnage des résultats	**OUI**	NON	OUI (parfois)
Extraction des informations derrière les liens	**OUI**	NON	NON
Gestion des formats bureautiques les plus répandus (Office, PDF,…)	**OUI**	OUI (parfois)	NON

*Comparatif donné à titre indicatif et basé sur le test des services indiqués ou sur leur documentation technique.

LES AVANTAGES DE L'APPROCHE DE DIGIMIND POUR LA SURVEILLANCE DU WEB INVISIBLE

Une recherche temps réel et sur mesure du Web Invisible

La possibilité d'ajouter ses propres moteurs à ses recherches et surveillances permet au veilleur d'effectuer une veille vraiment ciblée sur ses problématiques clés, évitant ainsi le bruit issu soit de moteurs non pertinents (pour les métamoteurs classiques du marché), soit de modifications non pertinentes de pages surveillées (pour les logiciels classiques de surveillance du web).

Il peut ainsi lancer des recherches sur des moteurs généralistes tels que *Yahoo!Search*, des moteurs spécialisés comme *Scirus* (recherche scientifique) ou des moteurs internes de bases de données, qu'elles soient médicales, juridiques ou scientifiques (*USPTO, PubMed, GlobalSpec*), et récolter immédiatement des résultats extrêmement pertinents, qu'il pourra ensuite mettre en surveillance en un seul clic.

Pour l'expert, mais aussi pour tout collaborateur de l'entreprise, c'est la possibilité de puiser en temps réel dans des ressources autrement fastidieuses à exploiter une par une – et de pouvoir répondre de manière beaucoup plus réactive et pertinente aux questions « urgentes » de sa hierarchie ou de ses collègues.

Une surveillance profonde du Web Invisible

La possibilité d'explorer en profondeur les documents vers lesquels pointent les résultats assure de plus de ne pas passer à côté d'une information clé qui n'aurait pas été indexée par un moteur classique.

Moins important peut-être dans une logique d'état de l'art sur un domaine, où les documents les plus représentatifs d'un sujet sont recherchés en priorité, et contiennent donc forcément les principaux mots-clés recherchés, l'exploration intégrale des documents devient vitale lorsqu'il s'agit d'identifier les « signaux faibles », par nature plus parcellaires, ambigus, et incomplets (*Lesca – Caron, 1986*).

L'avantage clé pour le veilleur ou l'expert de l'entreprise est une veille beaucoup plus fine et pertinente que celle qu'il pourrait effectuer avec des solutions "généralistes" du marché.

© Digimind

Une industrialisation de la surveillance

Cette veille plus pertinente débouche automatiquement sur un gain de temps important pour le veilleur : plus besoin de consulter des centaines de résultats – ou d'alertes – pour n'en retenir qu'une fraction.

Cette productivité est renforcée par une automatisation de l'intégralité du processus de recherche et surveillance du Web Invisible : application de requêtes à des bouquets de sources, surveillance profonde quels que soient les formats, langues ou encoding des sources, extraction et dédoublonnage automatiques des résultats, génération de résumés automatiques.

Enfin l'architecture J2EE et les fortes capacités de traitement mis à disposition par le centre serveur de Digimind permet d'assurer une veille sur un nombre de sources, de moteurs et de pages illimités.

Au final, le retour sur investissement de l'automatisation de la surveillance de sources électroniques, et en particulier du Web Invisible, dépasse les 400% pour atteindre dans certains cas 1000% (voir à ce sujet le White Paper de Digimind sur « *Evaluer le ROI d'un logiciel de veille* », par Edouard Fillias, Consultant Veille Stratégique chez Digimind).

Une démocratisation de la recherche et de la surveillance du web invisible

De part sa simplicité de paramétrage, le module de recherche et de surveillance du web invisible est accessible à des « non-spécialistes » des outils et de l'informatique, et en particulier aux experts métiers de l'entreprise (documentation, marketing, R&D, Produits, Finance,…).

La distribution de capacités de surveillance du web invisible dans l'entreprise à tous les experts métiers, qui vont pouvoir surveiller leurs propres sources, avec leur propre vocabulaire, et analyser les résultats avec leur expertise, enrichit fortement la qualité totale de la veille de l'entreprise.

Une logique de simplicité et de puissance que l'on retrouve dans tous les modules de la plateforme Digimind Evolution, et qui permet de déployer rapidement des projets de veille sur plusieurs dizaines et centaines de veilleurs.

© Digimind

A propos des auteurs

DIGIMIND CONSULTING

Digimind Consulting accompagne les plus grandes entreprises pour la conception, la mise en œuvre et le déploiement de projets de veille qui reposent sur la solution Digimind Evolution, apportant à ses clients un retour sur investissement de plus de 600%, et ce, dès la première année. Les méthodologies propriétaires du département conseil ainsi que son expertise des problématiques et sources d'informations sur plus de 10 secteurs d'activité, développée auprès de ses clients depuis de nombreuses années, permet aux entreprises d'anticiper les changements de leur environnement pour prendre les meilleures décisions sur leur marché.

Le conseil Digimind porte sur tous les aspects du workflow de veille : ciblage stratégique, collecte des informations, traitement et analyse, exploitation et diffusion, gestion de projet, conseil organisationnel et gestion du changement, formations et support fonctionnel et technique.

CHRISTOPHE ASSELIN

Issu d'un cabinet d'études de marché B2B et spécialisé depuis 1997 dans la mise en place de systèmes de veille (e-France.org, Ecole Militaire), Christophe Asselin allie une connaissance approfondie des secteurs économiques et une parfaite maîtrise des outils de recherche sur internet et des solutions avancées de veille. Expert reconnu, il édite le site http://www.intelligence-center.com et publie dans Netchercheur.

Consultant en veille stratégique chez Digimind, il accompagne les clients de Digimind dans la mise en place de leur dispositif de veille (expression des besoins, définition de plans de veille, sourcing,, architecture, paramétrage, formation, accompagnement). Il intervient ainsi auprès de sociétés dans différents secteurs : les télécoms avec Proximus, l'industrie pharmaceutique avec Sanofi Aventis, Roche Pharma, Expanscience ainsi que pour des références confidentielles dans le secteur des biotechnologies, l'industrie des transports, de l'imprimerie, de l'environnement, de la défense…

© Digimind

Notes

[1] http://www.google.com, http://www.google.fr, http://search.msn.com, http://search.msn.fr, http://fr.search.yahoo.com/, http://search.yahoo.fr

[2] http://search.yahoo.com/dir

[3] http://us.imdb.com

[4] URL : le terme URL (acronyme de Uniform Resource Locator address) désigne les adresses des documents internet. Pour le web, elles prendront la forme : http://www.site.fr

[5] http://www.alltheweb.com. Ce moteur norvégien a été racheté par *Overture* en février 2003. Overture a elle-même été rachetée par Yahoo! Inc. en juillet 2003. Aujourd'hui, l'index du moteur AlltheWeb est identique à celui de Yahoo!.

[6] http://beta.exalead.com/search, http://www.gigablast.com/

[7] "*Graph structure in the web*" http://www.almaden.ibm.com/cs/k53/www9.final , juin 2000. Andrei Broder[1], Ravi Kumar[2], Farzin Maghoul[1],
Prabhakar Raghavan[2], Sridhar Rajagopalan[2], Raymie Stata[3],
Andrew Tomkins[2], Janet Wiener[3]
1: AltaVista Company, San Mateo, CA.
2: IBM Almaden Research Center, San Jose, CA.
3: Compaq Systems Research Center, Palo Alto, CA.

[8] "*The Invisible Web: Finding Hidden Internet Resources Search Engines Can't See*"
Auteurs : Chris Sherman et Gary Price. 402 pages. Septembre 2001

[9] "*The Deep Web: Surfacing Hidden Value*" [9], Michael K. Bergman juillet 2000-février 2001
http://www.brightplanet.com/technology/deepweb.asp

[10] "*Searching the world Wide Web*" : Science Magazine, 3 avril 1998
http://clgiles.ist.psu.edu/papers/Science-98.pdf

[11] "*Accessibility of information on the web*" : Nature, 8 juillet 1999
(http://clgiles.ist.psu.edu/papers/Nature-99.pdf)

[12] http://www.inktomi.com/index.html

[13] Cyveillance, "*Sizing the Internet*" Brian H. Murray, 10 juillet 2000, Cyveillance,
http://www.cyveillance.com/web/downloads/Sizing_the_Internet.pdf

[14] IEEE Spectrum," *A Fountain of Knowledge*"; décembre 2003,
http://www.spectrum.ieee.org/WEBONLY/publicfeature/jan04/0104comp1.html

[15] http://ask.com

[16] http://dogpile.com
"*Missing Pieces*", 2005, par le métamoteur Dogpile.com en collaboration avec les chercheurs de University of Pittsburgh and The Pennsylvania State University.

"*Different Engines, Different Results. Web Searchers Not Always Finding What They're Looking for Online*"

[17] Jux2
http://www.jux2.com
http://www.jux2.com/about_different.php

[18] Chiffres Baromètre Xiti-1ère Position avril 2005 http://barometre.secrets2moteurs.com

[19] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

[20] Esp@ceNet http://ep.espacenet.com

[21] GlobalSpec http://www.globalspec.com

[22] Base de connaissance Microsoft http://support.microsoft.com/search/?adv=1

[23] Findarticles http://www.findarticles.com

[24] PlasticWay http://www.plasticway.com

[25] OCLC : http://www.oclc.org/, BNF-Gallica http://gallica.bnf.fr/, Library of Congress http://catalog.loc.gov/

[26] http://www.yellowpagesinc.com/, http://www.zoominfo.com/

[27] Traducteur Systran, http://babelfish.altavista.com/, Grand Dictionnaire Terminologique http://www.granddictionnaire.com

[28] http://www.chemicalonline.com, http://www.chemweb.com/, http://chemscope.com/, http://www.chem.com/, http://www.chemindustry.com/, http://www.france-chimie.com